WHAT IS CLAIMED IS:

1. A method of determining predictive models for a linked event detection system comprising the steps of:

determining source-identified training stories;

determining inter-story similarity vectors for at least one story-pair;

determining link label information for the at least one story-pair; and

determining at least one predictive model based on the inter-story similarity vector and the link label information.

2. The method of claim 1, wherein the step of determining inter-story similarity vectors comprises the steps of:

determining at least one inter-story similarity metric for the story-pairs; and

determining at least one source-pair statistics for the at least one story-pair.

3. The method of claim 2, wherein determining inter-story similarity vectors further comprise the step of normalizing the inter-story similarity metric based on the source-pair statistics.

4. The method of claim 2, wherein determining inter-story similarity vectors further comprise the step of incrementally normalizing the inter-story similarity metric based on the source-pair statistics.

5. The method of claim 2, wherein the inter-story similarity metric is normalized based on at least one of subtraction and division.

6. The method of claim 2, wherein the inter-story similarity metric is at least one of a probability based similarity metric and a Euclidean based similarity metric.

7. The method of claim 6, wherein the probability based inter-story similarity metric is at least one of a Hellinger, a Tanimoto and a clarity distance based metric.

8. The method of claim 6, wherein the Euclidean based inter-story similarity metric is a cosine-distance based metric.

9. The method of claim 1, further comprising the step of transforming the source-identified training stories.

10. The method of claim 9, wherein transforming the source-identified training stories is at least one of translating, transcribing and linguistically transforming.

11. The method of claim 1, wherein the inter-story similarity metrics are based on terms in at least one source-identified term frequency-inverse story frequency models.

12. The method of claim 11, wherein the terms in source-identified term frequency-inverse story frequency models are based on language.

13. The method of claim 11, wherein determining terms comprises the steps:

determining a reference language; and

determining reference language and non-reference language terms.

14. The method of claim 2, wherein the at least one inter-story similarity metric is normalized based on at least one of a source-pair identified similarity statistic.

15. The method of claim 1, wherein the at least one predictive model is at least one of: a classifier, a support vector machine, a decision tree and a Naive-Bayes classifier.

16. The method of claim 2, wherein at least one of the source-pair similarity statistics are determined based on a source hierarchy.

17. The method of claim 16 wherein the source hierarchy is determined based on at least one source characteristic.

18. The method of claim 16 wherein the source characteristic is at least one of a language characteristic, an input mode characteristic, a genre characteristic, a source name characteristic and a transformation characteristic.

19. The method of claim 16 wherein the source-pair similarity statistic for a new source is determined based on at least one source characteristic of the new source.

20. A linked event detection training system comprising:

an input/output circuit;

a memory;

a processor that receives source-identified training stories and associated link label information for at least one story-pair via the input/output circuit;

an inter-story similarity vector determining circuit that determines an inter-story similarity vector for at least one story-pair; and

a predictive model determining circuit that determines at least one predictive model based on the inter-story similarity vector and the link label information.

21. The system of claim 20, wherein the inter-story similarity vector determining circuit is comprised of:

a similarity metric determining circuit that determines at least one inter-story similarity metric for the at least one story-pair; and

a similarity statistics determining circuit that determines at least one source-pair statistic for the at least one story-pair.

22. The system of claim 21, wherein the inter-story similarity vector determining circuit normalizes the inter-story similarity metric based on the source-pair statistics.

23. The system of claim 21, wherein the inter-story similarity vector determining circuit incrementally normalizes the inter-story similarity metric based on the source-pair statistics.

24. The system of claim 21, wherein at least one of the inter-story similarity metrics is normalized based on at least one of a subtraction and a division operation.

25. The system of claim 21, wherein at least one of the inter-story similarity metrics is at least one of a probability based similarity metric and a Euclidean based similarity metric.

26. The system of claim 25, wherein the probability based inter-story similarity metric is at least one of a Hellinger, a Tanimoto and a clarity distance based metric.

27. The system of claim 25, wherein the Euclidean based inter-story similarity metric is a cosine-distance based metric.

28. The system of claim 20, wherein the source-identified training stories are transformed.

29. The system of claim 28, wherein transforming the source-identified training stories is at least one of translating, transcribing and linguistically transforming.

30. The system of claim 20, wherein the inter-story similarity metrics are based on terms in at least one source-identified term frequency-inverse story frequency model.

31. The system of claim 30, wherein the terms in the source-identified term frequency-inverse story frequency models are based on language.

32. The system of claim 30, wherein the processor determines terms based on a reference language; and determining reference language and non-reference language terms.

33. The system of claim 21 wherein the at least one inter-story similarity metric is normalized based on at least one of a source-pair identified similarity statistic.

34. The system of claim 20, wherein the at least one predictive model is at least one of: a classifier, a support vector machine, a decision tree and a Naive-Bayes classifier.

35. The system of claim 21, wherein the source-pair identified similarity statistic is determined based on a source hierarchy.

36. The system of claim 35, wherein the source hierarchy is determined based on at least one of a source characteristic.

37. The system of claim 35, wherein the source characteristic is at least one of a language characteristic, an input mode characteristic, a genre characteristic, a source name characteristic and a transformation characteristic.

38. The system of claim 35, wherein the source-pair similarity statistic for a new source is determined based on at least one source characteristics of the new source.

39. A method of linked event detection comprising the steps of:

determining source-identified training stories;

determining inter-story similarity vectors for the story-pairs;

determining at least one predictive model for link detection; and

determining a link between the story-pairs based on the predictive model and the inter-story similarity vector.

40. The method of claim 39, wherein the step of determining inter-story similarity vectors comprises the steps of:

determining at least one inter-story similarity metric for eachstory-pair; and

determining source-pair statistics for the story-pairs.

41. The method of claim 40, wherein determining inter-story similarity vectors further comprise the step of normalizing the inter-story similarity metric based on the source-pair statistics.

42. The method of claim 40, wherein determining inter-story similarity vectors further comprise the step of incrementally normalizing the inter-story similarity metric based on the source-pair statistics.

43. The method of claim 40, wherein the inter-story similarity metric is normalized based on at least one of subtraction and division.

44. The method of claim 40, wherein the inter-story similarity metric is at least one of a probability based similarity metric and a Euclidean based similarity metric.

45. The method of claim 44, wherein the probability based inter-story similarity metric is at least one of a Hellinger, a Tanimoto and a clarity distance based metric.

46. The method of claim 44, wherein the Euclidean based similarity metric is a cosine-distance based metric.

47. The method of claim 39, further comprising the step of transforming the source-identified training stories.

48. The method of claim 47, wherein transforming the source-identified training stories is at least one of translating, transcribing and linguistically transforming.

49. The method of claim 39, wherein the inter-story similarity metrics are based on terms in at least one source-identified term frequency-inverse story frequency models.

50. The method of claim 49, wherein the terms in source-identified term frequency-inverse story frequency models are based on language.

51. The method of claim 49, wherein determining terms comprises the steps:

determining a reference language; and

determining reference language and non-reference language terms.

52. The method of claim 40, wherein the at least one inter-story similarity metric is normalized based on at least one of a source-pair identified similarity statistic.

53. The method of claim 39, wherein the at least one predictive model is at least one of: a classifier, a support vector machine and a decision tree, a Naive-Bayes-classifier.

54. The method of claim 40, wherein the source-pair identified similarity statistic is determined based on a source hierarchy.

55. The method of claim 54, wherein the source hierarchy is determined based on at least one of a source characteristic.

56. The method of claim 54, wherein the source characteristic is at least one of a language characteristic, an input mode characteristic, a genre characteristic, a source name characteristic and a transformation characteristic.

57. The method of claim 54, wherein the source-pair similarity statistic for a new source is determined based on at least one source characteristics of the new source.

58. A linked event detection system comprising:

an input/output circuit;

a memory;

a processor that receives source-identified training stories via the input/output circuit;

an inter-story similarity vector determining circuit that determines inter-story similarity vectors for the story-pairs; and

a link determining circuit that determines links between story-pairs based on a predictive model and the inter-story similarity vectors.

59. The method of claim 58, wherein the inter-story similarity vector determining circuit is comprised of:

a similarity metric determining circuit that determines at least one inter-story similarity metric for the story-pairs; and

a similarity statistics determining circuit that determines source-pair statistics for the story-pairs.

60. The system of claim 59, wherein the inter-story similarity vector determining circuit normalizes the inter-story similarity metric based on the source-pair statistics.

61. The system of claim 59, wherein the inter-story similarity vector determining circuit incrementally normalizes the inter-story similarity metric based on the source-pair statistics.

62. The system of claim 59, wherein at least one of the inter-story similarity metrics is normalized based on at least one of a subtraction and a division operation.

63. The system of claim 59, wherein at least one of the inter-story similarity metrics is at least one of a probability based similarity metric and a Euclidean based similarity metric.

64. The system of claim 63, wherein the probability based inter-story similarity metric is at least one of a Hellinger, a Tanimoto and a clarity distance based metric.

65. The system of claim 63, wherein the Euclidean based inter-story similarity metric is a cosine-distance based metric.

66. The system of claim 58, wherein the source-identified training stories are transformed.

67. The system of claim 66, wherein transforming the source-identified training stories is at least one of translating, transcribing and linguistically transforming.

68. The system of claim 58, wherein the inter-story similarity metrics are based on terms in at least one source-identified term frequency-inverse story frequency model.

69. The system of claim 68, wherein the terms in the source-identified term frequency-inverse story frequency models are based on language.

70. The system of claim 68, wherein the processor determines terms based on a reference language; and non-reference language terms.

71. The system of claim 59, wherein the at least one inter-story similarity metric is normalized based on at least one of a source-pair identified similarity statistic.

72. The system of claim 58, wherein the predictive model is at least one of: a classifier, a support vector machine and a decision tree, a Naive-Bayes classifier.

73. The system of claim 59, wherein the source-pair identified similarity statistic is determined based on a source hierarchy.

74. The system of claim 73, wherein the source hierarchy is determined based on at least one of a source characteristic.

75. The system of claim 73, wherein the source characteristic is at least one of a language characteristic, an input mode characteristic, a genre characteristic, a source name characteristic and a transformation characteristic.

76. The system of claim 73, wherein the source-pair similarity statistic for a new source is determined based on at least one source characteristics of the new source.

77. A method of determining a stopword list comprising the steps of:

determining a source-identified training corpus of text information;

determining a verified first transformation of the source-identified training corpus text from a first source mode to a second source mode;

determining an un-verified second transformation of the source-identified training corpus text from a first source mode to a second source mode;

determining at least one transformation errors associated with distribution differences between the first and second transformations and identified sources;

determining at least one source-specific transformation actions for the determined transformation errors.

78. The method of claim 77, wherein the first source mode is at least one of a text source, an optical character recognition source and an automatic speech recognition source.

79. The method of claim 77, wherein the second source mode is at least one of a text source, an optical character recognition source and an automatic speech recognition source.

80. The method of claim 77, wherein the source-specific transformation is at least one of a removal, a repair and a normalization transformation.

81. Computer readable storage medium comprising: computer readable program code embodied on the computer readable storage medium, the computer readable program code usable to program a computer to determine at least one predictive model for a linked event detection system comprising the steps of:

determining source-identified training stories;

determining inter-story similarity vectors for at least one story-pair;

determining link label information for the at least one story-pair; and

determining at least one predictive model based on the inter-story similarity vector and the link label information.

82. A carrier wave encoded to transmit a control program, useable to program a computer to determine a predictive model for a linked event detection system, to a device for executing the program, the control program comprising:

instructions for determining source-identified training stories;

instructions for determining inter-story similarity vectors for at least one story-pair;

instructions for determining link label information for the at least one story-pair; and

instructions for determining at least one predictive model based on the inter-story similarity vector and the link label information.

83. Computer readable storage medium comprising: computer readable program code embodied on the computer readable storage medium, the computer readable program code usable to program a computer to detect linked events comprising the steps of:

determining source-identified training stories;

determining inter-story similarity vectors for the the at least one story-pair;

determining at least one predictive model for link detection;

determining a link between story-pairs based on the at least one predictive model and the inter-story similarity vector.

84. A carrier wave encoded to transmit a control program, useable to program a computer to detect linked events, to a device for executing the program, the control program comprising:

instructions for determining source-identified training stories;

instructions for determining inter-story similarity vectors for the at least one story-pair;

instructions for determining at least one predictive model for link detection;

instructions for determining a link between story-pairs based on the predictive model and the inter-story similarity vector.

85. The method of claim 2, wherein determining at least one source-pair statistic for the at least one story-pair is based on at least one of a similarity metric and a statistic associated with the metric.

86. The system of claim 21, wherein determining at least one source-pair statistic for the at least one story-pair is based on at least one of a similarity metric and a statistic associated with the metric.

87. The method of claim 39, wherein at least one of the predictive models is a trained predictive model.

88. The system of claim 58, wherein at least one of the predictive models is a trained predictive model.